WARWICK
THE UNIVERSITY OF WARWICK

Original citation:
McCaig, Duncan, Bhatia, Sudeep, Elliott, Mark T., Walasek, Lukasz and Meyer, Caroline (2018) Text-mining as a methodology to assess eating disorder-relevant factors : comparing mentions of fitness tracking technology across online communities. International Journal of Eating Disorders . doi:10.1002/eat.22882

Permanent WRAP URL:
http://wrap.warwick.ac.uk/102764

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
This is the peer reviewed version of the following article: McCaig D, Bhatia S, Elliott MT, Walasek L, Meyer C. Text-mining as a methodology to assess eating disorder-relevant factors: Comparing mentions of fitness tracking technology across online communities. Int J Eat Disord. 2018. https://doi.org/10.1002/eat.22882 , which has been published in final form at https://doi.org/10.1002/eat.22882 . This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

A note on versions:
The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP url' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications

Title:

Text-mining as a methodology to assess eating disorder-relevant factors: Comparing mentions of fitness tracking technology across online communities

Running title:

Text-mining and eating disorders

Authorship:

Duncan McCaig[a,*], Sudeep Bhatia[b], Mark T. Elliott[a], Lukasz Walasek[a] and Caroline Meyer[a,c,d]

*Corresponding author. *Email address*: Duncan McCaig (d.mccaig@warwick.ac.uk)

Authors' institutional affiliations:

[a]WMG, University of Warwick, Coventry, UK

[b]Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

[c]Warwick Medical School, University of Warwick, Coventry, UK

[d]University Hospitals Coventry and Warwickshire NHS Trust, UK

Acknowledgements:

Duncan McCaig is the recipient of a doctoral studentship from WMG, University of Warwick, and Coventry and Warwickshire NHS Partnership Trust. Mark Elliott is currently funded by Innovate UK (Project ref: 80753-504440).

Conflict of interest statement:

Mark Elliott is currently undertaking a collaborative research project with Sweatco Ltd, producers of the Sweatcoin app. This project is funded by Innovate UK. Sweatco Ltd has had no involvement in the research presented in this report at any stage. The remaining authors declare no conflict of interest.

Abstract

Objective: Text-mining offers a technique to identify and extract information from a large corpus of textual data. As an example, this study presents the application of text-mining to assess and compare interest in fitness tracking technology across eating disorder and health-related online communities. *Method*: A list of fitness tracking technology terms was developed, and communities (i.e., 'subreddits') on a large online discussion platform (*Reddit*) were compared regarding the frequency with which these terms occurred. The corpus used in this study comprised all comments posted between May 2015 and January 2018 (inclusive) on six subreddits – three eating disorder-related, and three relating to either fitness, weight-management or nutrition. All comments relating to the same 'thread' (i.e., conversation) were concatenated, and formed the cases used in this study (*N*=377,276). *Results*: Within the eating disorder-related subreddits, the findings indicated that a 'pro-eating disorder' subreddit, which is less recovery focused than the other eating disorder subreddits, had the highest frequency of fitness tracker terms. Across all subreddits, the weight-management subreddit had the highest frequency of the fitness tracker terms' occurrence, and *MyFitnessPal* was the most frequently mentioned fitness tracker. *Discussion*: The technique exemplified here can potentially be used to assess group differences to identify at-risk populations, generate and explore clinically relevant research questions in populations who are difficult to recruit, and scope an area for which there is little extant literature. The technique also facilitates methodological triangulation of research findings obtained through more 'traditional' techniques, such as surveys or interviews.

Keywords: eating disorders, fitness tracking, mental health, social media, text-mining

1. Introduction

Text-mining enables the identification and extraction of specific information from a large corpus (i.e., collection) of textual data. This technique could provide a wealth of insight for the field of eating disorders, and has previously been used in various ways within psychiatry (Abbe, Grouin, Zweigenbaum, & Falissard, 2016).

As one of the largest online discussion platforms, *Reddit* (https://www.reddit.com/) offers a vast source of publicly available text on which text-mining can be performed. Within *Reddit*, there are 'subreddits' that relate to different topics (e.g., politics, films). These subreddits can be viewed as communities of people with shared interests. Users can post on a subreddit to start a discussion, then other people can comment on the post. This creates a series of comments in response to a post (i.e., a 'thread'), which can be conceptualised as a conversation.

As several subreddits relate specifically to eating disorders, the textual data available through *Reddit* offers the potential to improve our understanding of these communities. Several studies have qualitatively analysed the content of online eating disorder communities (e.g., Borzekowski, Schenk, Wilson, & Peebles, 2010; Branley & Covey, 2017; Sowles et al., 2018; Teufel et al., 2013). This content can be broadly categorised as either: 'pro-eating disorder', which depicts a desire to enact eating disorder behaviours without indicating a desire to recover; or 'pro-recovery' (or 'anti-eating disorder'), which encourages recovery from eating disorders and/or resistance to eating disorder cognitions and behaviours (cf. Branley & Covey, 2017). Pro-eating disorder communities have been indicated to be the most active, as these groups were observed to post more frequently than pro-recovery groups (Teufel et al., 2013). Regarding individual characteristics, large cross-sectional surveys have found that more frequent interaction with pro-eating disorder online content is associated with higher levels of eating disorder symptoms, and a higher lifetime occurrence of binge eating, purging, laxative use and diet pill use (Harper, Sperry, & Thompson, 2008; Peebles et al., 2012). In Peebles and colleagues' study surveying pro-eating disorder website

users (Peebles et al., 2012), and a separate study surveying pro-recovery website users (Aardoom, Dingemans, Boogaard, & Van Furth, 2014), both communities were found to have relatively high levels of eating disorder symptomatology, with the mean score of pro-eating disorder website users in Peebles and colleagues' study exceeding the typical clinical cut-off (e.g., Carter, Stewart, & Fairburn, 2001). Due to the differences between these online communities in terms of the extent to which they focus on recovery, the content posted in these communities can facilitate the investigation of questions regarding eating disorder readiness to change (cf. Geller, Srikameswaran, Brown, Piper, & Dunn, 2013).

In areas in which there is little extant research, text-mining can be particularly beneficial for scoping a research question. One such area is used to exemplify this technique – fitness tracking technology and its relationship with eating disorder symptomatology. Fitness tracking technology (i.e., fitness trackers) can be defined as devices that enable the self-monitoring of one's physical activity, such as calorie-expenditure (Simpson & Mazzeo, 2017). Recently, studies have investigated the nature of fitness tracker usage in relation to the eating disorders (Hefner et al., 2016; Levinson, Fewell, & Brosof, 2017; Simpson & Mazzeo, 2017; Tan, Kuek, Goh, Lee, & Kwok, 2016). All of these studies used self-report surveys and their findings are largely in agreement. In community samples, fitness tracker usage was found to be positively associated with eating disorder symptomatology (Hefner et al., 2016; Simpson & Mazzeo, 2017). In the context of eating disorder recovery, more clinical eating disorder patients were found to view fitness trackers as contributing to the maintenance of their eating disorder rather than aiding recovery (Tan et al., 2016). Similarly, a majority of recently discharged eating disorder patients were indicated to have used the app *MyFitnessPal*, and viewed it as having contributed to their eating disorder (Levinson et al., 2017).

The findings from the previously detailed survey-based studies are subject to two main limitations. First, when assessing fitness tracker usage, the majority of studies used dichotomous response scales (i.e., responding 'yes' or 'no'; Levinson et al., 2017; Simpson & Mazzeo, 2017; Tan et

al., 2016). Dichotomous response scales are problematic as they might increase Type I (false positive) error rate, underestimate variation in the sample, and conceal non-linearity (Altman & Royston, 2006; Austin & Brunner, 2004). These problems might cause researchers to conclude that there is a relationship when there is actually insufficient evidence. Alternatively, researchers might fail to identify a more complex relationship. Second, survey-based methodologies might have inadvertently influenced participants to give an affirmative response (e.g., "Did you feel that My Fitness Pal contributed to your eating disorder in any way?"; Levinson et al., 2017).

Through the application of text-mining to compare the frequency with which fitness trackers are mentioned across eating disorder subreddits, the two limitations of the previous research can be addressed. First, as frequency counts obtained through text-mining provide continuous data, no such dichotomising of measurement is undertaken. Second, as text-mining is a primarily data-driven approach, there is less risk of researchers inadvertently biasing findings. As text-mining can address these limitations, its strength is as a complementary technique, and as a way to methodologically triangulate findings obtained through more 'traditional' techniques (e.g., surveys). While 'mentions' of fitness trackers cannot be assumed to reflect their usage, they can be interpreted as an indication of interest (e.g., Walasek, Bhatia, & Brown, 2017).

In summary, recent research concerning the relationship between eating disorder symptomatology and fitness tracker usage has two main limitations – dichotomising measurement, and potentially having influenced participants' responses. The primary aim of the current research was to address these limitations by using text-mining to explore how pro-eating disorder and pro-recovery subreddits differed in the frequency with which fitness trackers were mentioned. A secondary aim of the research was to identify whether fitness trackers were most frequently mentioned in eating, body shape and weight, or exercise-related contexts. More generally, the current research aimed to exemplify the application of text-mining to explore an eating disorder-

relevant research question. As this study was exploratory in nature, no *a priori* hypotheses were stated.

2. Methods

2.1. Overview of method

A flowchart summarising the procedure is presented in Figure 1. Following selection of the corpus and generation of terms, data of interest were extracted and pre-processed, and threads were concatenated. The data were then analysed, which included summary of the corpus' characteristics, and comparison of the subreddits regarding the frequency with which different terms were mentioned.

[insert Fig. 1 here]

2.2. Corpus selection

All public *Reddit* comments since December 2005 are freely available from a regularly updated archive (Complete Public Reddit Comments Corpus, 2015). This archive does not include the initial post to which commenters responded. The entire archive between May 2015 and January 2018 (inclusive) was downloaded. As fitness trackers' ubiquity and popularity vary across time, it was deemed important that all of the subreddits covered the same time period. May 2015 was selected as the start point, as this represented the earliest month at which all identified subreddits were active. Descriptions of the included subreddits are presented in Table 1. As the analyses in this study were restricted to publicly available data, an exemption from ethical review was obtained for this study from the University of Warwick's Biomedical and Scientific Research Ethics Committee.

2.2.1. Eating disorder subreddits. Reddit's search bar was used to identify eating disorder-related subreddits that contained one or more eating disorder-related term in the subreddit's name and/or description. The list of search terms was generated through consultation of two clinical references

(DSM-V; American Psychiatric Association, 2000; ICD-10; World Health Organization, 1993), and previous research (e.g., Branley & Covey, 2017; Chancellor, Lin, Goodman, Zerwas, & De Choudhury, 2016). The following search terms were developed that related to eating disorders in general: "eating disorder", "eating disorders", "eating disordered", "eatingdisorder", "eatingdisorders", "eatingdisordered", "disordered eating", "disorderedeating", "ed", "eds", "proed", "proeds", "pro-ed" and "pro-eds". Due to the focus on fitness trackers, the following search terms were also developed for eating disorder diagnostic categories that included criteria relating to exercise (i.e., Anorexia Nervosa and Bulimia Nervosa): "anorexia", "anorexic", "anorexics", "proanorexia", "pro-anorexia", "ana", "proana", "pro-ana", "bulimia", "bulimic", "bulimics", "probulimia", "pro-bulimia", "mia", "promia" and "pro-mia".

From the resultant list of eating disorder subreddits, the three with the most threads (see section 2.5.) were selected for inclusion. These were *r/proED, r/fuckeatingdisorders* and *r/EatingDisorders*.

2.2.2. Health-related subreddits. Three large subreddits (i.e., >150,000 subscribers on December 31 2017) were also included in the corpus to explore the context in which the fitness trackers were most frequently mentioned. Three subreddits were selected that each related to one of three eating disorder-related behaviours, or behavioural outcomes – eating (i.e., *r/nutrition*), body shape and weight (i.e., *r/loseit*, a weight-management subreddit), and exercise (i.e., *r/Fitness*).

[insert Table 1 here]

2.3. Generation of terms

Complete lists of the terms detailed below are provided as Supporting Information. In line with the data pre-processing approach (section 2.4., step 1), all terms are lowercase.

2.3.1. Fitness tracker terms. A list of nouns relating to fitness trackers was developed by consulting a comprehensive website of fitness wearables (inKin Social Fitness, 2017), Google Play 'Get Fighting

Fit' and 'Get Outside' health and fitness app categories (Google, 2017), and previous literature relating to fitness trackers and eating disorders (e.g., Levinson et al., 2017). Generic terms (e.g., "fitness tracker") and other fitness trackers of theoretical interest (e.g., "cronometer") were also added.

Once this list had been developed, all multiword terms were added to the list of fitness tracker terms without whitespace (e.g., "fitbitsurge"). For single-word terms, each term was first entered separately into an internet search engine. In the case of multiword terms that corresponded to a brand/make (e.g., "fitbit") and model/app (e.g., "surge"), each term was also entered separately. For each separate term, if one or more of the top three search results related to the fitness tracker, it was deemed to have sufficient brand presence to be added to the list of fitness tracker terms on its own. Once the list of terms was compiled, any numeric values in the terms were removed (e.g., "couch25k" was translated into "couchk")

In all subsequent analyses, exact matches of the fitness tracker terms were sought. Therefore, as the terms represent nouns, singular and plural forms of each term were generated. First, the list of terms was reviewed, and irregular plurals – i.e., not created by only appending an "s" or "es" suffix to the singular form – were created on a case-by-case basis, and added as an additional term. For example, "bellabeatleaf" was pluralised to "bellabeatleaves". Two variations of each term were then generated, which represented two plural forms comprising the suffixes 's' and 'es' appended to the singular form (e.g., "fitbits"). This approach resulted in plurals that were not necessarily correct (e.g., "fitbites"). Despite this, the approach was undertaken as it is more reproducible than manually reviewing the term list and removing any ostensibly incorrect plurals. In addition, commenters might not necessarily pluralise nouns correctly (e.g., "apple watchs"). As a result, the liberal approach used here also identified commenters' incorrect plurals, which were viewed as being of equal semantic importance as correct plurals.

The resultant list included 169 unique fitness tracker terms. As detailed above, each fitness tracker term also had two additional plural forms (i.e., +"s" and +"es"), resulting in a total of 507 terms.

2.3.2. Recovery, eating, body and exercise-related terms. Three separate lists of terms were developed that related to either eating, the body or exercise. These terms were generated by compiling a list of all related terms taken from validated self-report measures. Eating terms (e.g., "calories") were generated from the Eating Disorder Examination Questionnaire (EDE-Q; Fairburn & Beglin, 2008) Restraint and Eating Concerns subscales, and the Shape and Weight Concerns subscales were used for body terms (e.g., "weight"). Exercise terms (e.g., "run") were generated from the International Physical Activity Questionnaire (Craig et al., 2003) and Compulsive Exercise Test (Taranis, Touyz, & Meyer, 2011). For each identified term, several word forms were also generated (e.g., "exercise", "exercises", "exercised", "exercising"). A list of five recovery terms was also generated comprising the term "recovery" and four related word forms.

2.4. Data extraction and pre-processing

The procedure undertaken to process the corpus is detailed in the following steps:

1) For all the comments within the corpus, the following information was extracted: month posted, author identifier, subreddit name, thread identifier, and comment text. At the point of extraction, each unique commenter was assigned an author identifier number, which was saved in place of their username. Any comments created by the automated moderation bot ("AutoModerator") were excluded. To avoid variations in capitalisation in the corpus, each comment's text was translated into lowercase characters. The comment text was then pre-processed by removing the following: URL links, the phrase "[deleted]" (representing a comment deleted before archiving), punctuation, numeric characters, and common English stopwords, such as personal pronouns (e.g., "my"). In each comment's text, all occurrences

of multiword fitness tracker terms (section 2.3.) were also concatenated by removing whitespace (e.g., “fitbit surge” was translated into “fitbitsurge”).

2) The text of all comments that had the same subreddit name and thread identifier was then concatenated, which produced the corpus of pre-processed threads. Threads were selected rather than comments, as all comments in a thread correspond to the same initial post. Therefore, comments cannot be deemed to be independent, whereas threads can.

All data pre-processing was undertaken using the freely available natural language toolkit (Bird, Klein, & Loper, 2009), and all code was written in Python programming language (Python Software Foundation, 2017). The code relating to the two steps detailed above is available in an online repository (https://github.com/mccaigduncan/Text-mining-and-eating-disorders), and can be used to replicate the data extraction and pre-processing steps with the same or different subreddits.

2.5. Data analysis

The following steps were used to analyse the data:

1) Characteristics of the corpus and the subreddits were calculated (e.g., number of threads and commenters).
2) For the eating disorder subreddits, each thread containing at least one recovery term was counted, and the percentage of threads within each subreddit that referenced recovery was calculated. This method was repeated for each of the eating, body and exercise lists of terms for all six subreddits in the corpus.
3) The percentage of threads referencing fitness trackers was calculated for each of the six subreddits using the same method as detailed in step 2). The same process was then used to calculate the percentage of threads within which each individual fitness tracker term occurred. After identifying the three most frequently mentioned fitness trackers across all six subreddits, all terms that occurred within the corpus that related to each specific tracker

were grouped (e.g., “mfp” and “myfitnesspal” for *MyFitnessPal*). Using the same method as in step 2), these groups of terms were used to calculate the percentage of threads that referenced each fitness tracker.

3. Results

3.1. Corpus characteristics

For each of the six subreddits, descriptive statistics regarding the number of threads, comments and unique commenters, and the average number of comments made by each unique commenter are presented below in Table 2.

[insert Table 2 here]

Across all six subreddits, there were a total of 377,276 threads, 7,044,686 comments and 508,742 unique commenters. Each unique commenter posted on an average of one out of the six subreddits (*M*=1.08, *SD*=.30; median=1, range=1:6).

3.2. Mentions of recovery in eating disorder subreddits

9.75% of threads within the *r/proED* subreddit mentioned recovery, compared to 43.17% of *r/fuckeatingdisorders* threads and 50.41% of *r/EatingDisorders* threads.

3.3. Mentions of eating, the body and exercise in all subreddits

Figure 2 presents separately the percentage of each subreddits’ threads that mentioned at least one term from each of the eating, body and exercise-related lists of terms.

[insert Fig. 2 here]

Figure 2 indicates that eating was most frequently mentioned in *r/nutrition*, the body was most frequently mentioned in *r/loseit*, and exercise was most frequently mentioned in *r/Fitness*.

r/loseit was also indicated to have the second highest proportion of threads mentioning both eating and exercise-related terms.

3.4. Mentions of fitness trackers in all subreddits

Figure 3 presents the percentage of threads within a subreddit that contained one or more of the fitness tracker terms.

[insert Fig. 3 here]

Regarding the eating disorder subreddits, Figure 3 indicates that 5.49% of threads within the *r/proED* subreddit contained a reference to a fitness tracker, which was greater than the other eating disorder subreddits, *r/fuckeatingdisorders* and *r/EatingDisorders* (1.88% and 2.18%, respectively). Regarding the health-related subreddits, *r/loseit* had a higher percentage (30.65%) than *r/Fitness* and *r/nutrition* (5.65% and 9.05%, respectively).

3.5. Frequently mentioned fitness trackers

The three most frequently mentioned fitness trackers within the corpus were identified. All of the terms that related to each identified tracker and occurred in the corpus were grouped. Accordingly, the three identified fitness trackers were *MyFitnessPal*, *Fitbit* and *Heart rate monitor*. The terms relating to each fitness tracker are provided as Supporting Information, and the data for the frequency of mentions for each specific term are available in the online repository detailed in section 2.4. Figure 4 presents the percentage of threads identified in section 3.4. that included at least one reference to each aforementioned fitness tracker.

[insert Fig. 4 here]

As shown in Figure 4, among the threads in which fitness trackers were mentioned, more than 40% included mentions of *MyFitnessPal*. This was true for all six subreddits, although such mentions were particularly prevalent in *r/loseit*.

4. Discussion

This study aimed to apply text-mining to online communication to investigate an eating disorder-relevant research question. The application of this technique was exemplified by analysing the relative frequencies with which fitness tracker terms were mentioned within different online communities (subreddits). Within the eating disorder subreddits, the fitness tracker terms were most frequently mentioned in the least recovery-focused subreddit. Within health-related subreddits, the highest proportion of mentions was in the weight-management subreddit. Regarding specific fitness trackers, *MyFitnessPal* was the most frequently mentioned in all subreddits, and occurred in 40% or more of the threads that mentioned fitness trackers. A strength of these findings is that they were obtained through the unsolicited communication of over half a million users, and were therefore not influenced by biases that are common in typical quantitative and qualitative methodologies.

Regarding eating disorder subreddits, fitness trackers were more frequently mentioned in *r/proED*, than in *r/fuckeatingdisorders* or *r/EatingDisorders*. Approximately ten percent of *r/proED* threads mentioned recovery, compared to approximately half of the threads in *r/fuckeatingdisorders* and *r/EatingDisorders*. This suggests that *r/proED* is less recovery-focused than *r/fuckeatingdisorders* and *r/EatingDisorders*. Due to the more frequent mentions of fitness trackers in *r/proED*, this finding suggests that, while self-reported fitness tracker usage has been positively associated with eating disorder symptomatology (Hefner et al., 2016; Simpson & Mazzeo, 2017), this relationship might be more nuanced. Rather, the association between fitness tracker usage and these symptoms might be moderated by a person's stage of change regarding recovery (cf. Geller et al., 2013). As fitness trackers were more frequently mentioned in the least recovery-focused subreddit (*r/proED*), the current findings support previous survey-based research that indicated fitness tracker usage to be

more associated with the maintenance of eating disorders than recovery from them (Levinson et al., 2017; Tan et al., 2016).

Within the health-related subreddits, fitness tracker terms were most frequently mentioned in the weight-management subreddit (*r/loseit*), which can be interpreted as a higher interest in fitness trackers in this community (e.g., Walasek et al., 2017). As *r/loseit* was indicated to be the health-related subreddit with the most frequent mentions of body terms, the findings suggest that fitness tracker interest is particularly high in people with a high interest in the body. This is in line with previous findings that fitness tracker usage was positively associated with shape and weight concerns (Simpson & Mazzeo, 2017). Undertaking exercise for weight-management (i.e., 'weight control exercise') is a key dimension of compulsive exercise, and has been linked with shape and weight concerns in both community and clinical eating disorder samples (e.g., Noetel et al., 2016; Taranis et al., 2011). A positive association between overall compulsive exercise and usage of apps (including fitness trackers) has also been observed (Hefner et al., 2016). Taken together, these findings suggest that shape and weight concerns, weight control exercise and fitness tracker usage are likely to be inter-related. Overall, fitness trackers were more frequently mentioned in the health-related subreddits than in the eating disorder subreddits, which likely reflects a narrower content focus of the health-related subreddits.

The current study also found that *MyFitnessPal* was the most mentioned fitness tracker in each subreddit, which supports a research focus on the usage of this particular fitness tracker in an eating disorder population (cf. Levinson et al., 2017). Additionally, commenters in *r/proED* were shown to have a higher average number of comments than commenters in *r/fuckeatingdisorders* and *r/EatingDisorders*, which is in line with the finding that pro-eating disorder groups posted more frequently than pro-recovery groups (Teufel et al., 2013).

Overall, the example presented here supports the application of text-mining to complement 'traditional' methodologies, as the current findings converged with those previously obtained

through survey-based measures (Hefner et al., 2016; Levinson et al., 2017; Simpson & Mazzeo, 2017; Tan et al., 2016). However, both the text-mining and survey-based findings require further validation, as the data obtained through these methods might not represent actual fitness tracker usage.

There are several limitations of the text-mining approach. First, homonymy (i.e., similarly spelled words with different meanings) and polysemy (i.e., one word with several meanings) are problematic for text-mining (Abbe et al., 2016). For example, “apple” could refer to the fruit or the brand of fitness tracker. As such, if “apple” had been included in the current research, the term could have inflated frequency counts. In order to mitigate this, terms that did not identify a fitness tracker in an internet search were excluded. A second general limitation is that naturally occurring language might include typographic errors or variations in spelling. As such, semantically relevant terms might not be identified due to these errors. This limitation was minimised in the current research by generating correct and incorrect plurals. Third, the corpus used in text-mining might be subject to selection bias. For example, *Reddit* was selected as it represents a large publicly available source of data. However, users of the eating disorder subreddits might differ from those who visit professionally managed online forums. Alternatively, a different pattern of results might have been observed if a narrower time period had been selected. In order to overcome the effect of selection bias, replication of the analyses using different sources of data should be conducted. Finally, the generation of terms might introduce an element of subjectivity into text-mining. For example, in the current research, some fitness trackers might not have been identified. This subjectivity has been accounted for by ensuring that the search terms in this report were clearly detailed (enabling replication and extensions), and by making relative comparisons across subreddits.

Regarding the specific application of text-mining exemplified above, a limitation is that the included subreddits were assumed to represent people with specific characteristics (e.g., interested in weight-management). However, the actual characteristics of these commenters were unknown.

While this limits the conclusions in the current research, this is not an issue for text-mining as a technique, as it can be used with any large corpus of text (Abbe et al., 2016). As such, if researchers had access to a corpus of text for which they had more detail (e.g., demographics of each commenter), exactly the same technique could be undertaken. Future research could assess differences between the subreddits regarding the commenters' characteristics (e.g., levels of eating disorder symptomatology).

In comparison to qualitative techniques, text-mining is a rapid technique for identifying and extracting salient information, and can be used with any large corpus of text. As such, it offers several potentially beneficial applications for clinically relevant research. First, text-mining can facilitate the identification of groups that are potentially at-risk for eating disorders. As exemplified in the current research, an ostensibly pro-eating disorder community mentioned fitness trackers more frequently than pro-recovery communities. A similar application of this technique could be applied to investigate group differences in eating disorder diagnostic criteria, such as purging or laxative use. Through identifying group differences, more targeted approaches could be introduced (e.g., screening, interventions). Similarly, this technique facilitates the generation and exploration of clinically relevant research questions for which there is little extant literature, or in populations who might otherwise be difficult to recruit. For example, the pro-eating disorder population might be difficult to recruit due to being characterised as non-recovery focused (Branley & Covey, 2017).

The application of text-mining exemplified above could be enhanced to investigate other aspects of the textual data associated with clinically relevant variables. For example, all comments that mentioned fitness trackers could be identified and extracted using the code provided. The sentiment of these comments could then be assessed using automatic techniques (e.g., Thelwall, Buckley, & Paltoglou, 2012) and compared to comments with no mentions of fitness trackers. Alternatively, 'traditional' qualitative analyses (e.g., thematic analysis) could be undertaken on these comments to explore emergent themes.

From a research standpoint, a strength of text-mining is that it enables the triangulation of findings from studies that use different methodologies. As previously detailed, the current application supported findings obtained through quantitative survey-based methods (e.g., Simpson & Mazzeo, 2017). By exploring the same research question with various techniques, methodology-specific limitations can be mitigated. As exemplified above, as a data-driven approach, the application of text-mining in this study was not susceptible to the limitations of the previous survey-based studies (e.g., dichotomising measurement, potentially influencing responses). Similarly, the survey-based studies are less susceptible to the limitations of this application of text-mining (e.g., unassessed participant characteristics, homonymy/polysemy). As each methodology addresses limitations of the other, the findings obtained using both techniques will be less susceptible to biases than findings using only one. Text-mining could also be conducted in parallel with a literature review to address a research question, and generate hypotheses for confirmatory research. This would enable a more complete overview of current evidence, particularly as the relevance of extant research findings might have diminished if published a relatively long time ago.

In conclusion, text-mining can be used to identify illuminating patterns in an unstructured corpus of text. As exemplified in this study, the technique can be used for purposes such as triangulating findings obtained through different methodologies, scoping areas for which there is little extant research, identifying at-risk populations, and generating and exploring research questions in populations who are difficult to recruit. Due to these potentially great benefits, and as the technique is relatively rapid to undertake, it is argued that the application of text-mining is strongly warranted in the eating disorder field.

References

Aardoom, J. J., Dingemans, A. E., Boogaard, L. H., & Van Furth, E. F. (2014). Internet and patient empowerment in individuals with symptoms of an eating disorder: a cross-sectional investigation of a pro-recovery focused e-community. *Eat Behav, 15*(3), 350-356. doi:10.1016/j.eatbeh.2014.04.003

Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry: a systematic literature review. *Int J Methods Psychiatr Res, 25*(2), 86-100. doi:10.1002/mpr.1481

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ, 332*, 1080.

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR* (4th Ed.). Washington DC: American Psychiatric Association.

Austin, P. C., & Brunner, L. J. (2004). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Stat Med, 23*(7), 1159-1178. doi:10.1002/sim.1687

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*: O'Reilly Media.

Borzekowski, D. L., Schenk, S., Wilson, J. L., & Peebles, R. (2010). e-Ana and e-Mia: A content analysis of pro-eating disorder Web sites. *Am J Public Health, 100*(8), 1526-1534. doi:10.2105/AJPH.2009.172700

Branley, D. B., & Covey, J. (2017). Pro-ana versus Pro-recovery: A Content Analytic Comparison of Social Media Users' Communication about Eating Disorders on Twitter and Tumblr. *Frontiers in Psychology, 8*. doi:10.3389/fpsyg.2017.01356

Carter, J. C., Stewart, D. A., & Fairburn, C. G. (2001). Eating disorder examination questionnaire: Norms for young adolescent girls. *Behaviour Research and Therapy, 39*(5), 625-632.

Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). *Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities*. Paper

presented at the Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16.

Complete Public Reddit Comments Corpus. (2015). *Retrieved from https://archive.org/details/2015_reddit_comments_corpus.*

Craig, C. L., Marshall, A. L., Sjostrom, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., . . . Oja, P. (2003). International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc, 35*(8), 1381-1395. doi:10.1249/01.MSS.0000078924.61453.FB

Fairburn, C. G., & Beglin, S. J. (2008). Eating Disorder Examination Questionnaire (EDE-Q 6.0). In C. G. Fairburn (Ed.), *Cognitive behavior therapy and eating disorders.* (pp. 309-314). New York: Guilford Press.

Geller, J., Srikameswaran, S., Brown, K. E., Piper, W., & Dunn, E. C. (2013). The Psychometric Properties of the Readiness and Motivation Questionnaire: A Symptom-Specific Measure of Readiness for Change in the Eating Disorders. *Psychological Assessment, 25*(3), 759-768. doi:10.1037/a0032539.supp

Google. (2017). Health and Fitness Apps. *Retrieved from play.google.com/store/apps/category/HEALTH_AND_FITNESS?hl=en_GB*.

Harper, K., Sperry, S., & Thompson, J. K. (2008). Viewership of pro-eating disorder websites: association with body image and eating disturbances. *Int J Eat Disord, 41*(1), 92-95. doi:10.1002/eat.20408

Hefner, V., Dorros, S. M., Jourdain, N., Liu, C., Tortomasi, A., Greene, M. P., . . . Alvares, C. (2016). Mobile exercising and tweeting the pounds away: The use of digital applications and microblogging and their association with disordered eating and compulsive exercise. *Cogent Social Sciences, 2*(1). doi:10.1080/23311886.2016.1176304

inKin Social Fitness. (2017). Wearables. Retrieved from www.inkin.com/wearables/

Levinson, C. A., Fewell, L., & Brosof, L. C. (2017). My Fitness Pal calorie tracker usage in the eating disorders. *Eat Behav, 27*, 14-16. doi:10.1016/j.eatbeh.2017.08.003

Noetel, M., Miskovic-Wheatley, J., Crosby, R. D., Hay, P., Madden, S., & Touyz, S. (2016). A clinical profile of compulsive exercise in adolescent inpatients with anorexia nervosa. *J Eat Disord, 4*, 1. doi:10.1186/s40337-016-0090-6

Peebles, R., Wilson, J. L., Litt, I. F., Hardy, K. K., Lock, J. D., Mann, J. R., & Borzekowski, D. L. (2012). Disordered eating in a digital age: eating behaviors, health, and quality of life in users of websites with pro-eating disorder content. *J Med Internet Res, 14*(5), e148. doi:10.2196/jmir.2023

Python Software Foundation. (2017). Python Language Reference (Version 3.6.3.). Available at http://www.python.org.

Simpson, C. C., & Mazzeo, S. E. (2017). Calorie counting and fitness tracking technology: Associations with eating disorder symptomatology. *Eat Behav, 26*, 89-92. doi:10.1016/j.eatbeh.2017.02.002

Sowles, S. J., McLeary, M., Optican, A., Cahn, E., Krauss, M. J., Fitzsimmons-Craft, E. E., . . . Cavazos-Rehg, P. A. (2018). A content analysis of an online pro-eating disorder community on Reddit. *Body Image, 24*, 137-144. doi:10.1016/j.bodyim.2018.01.001

Tan, T. N., Kuek, A., Goh, S. E., Lee, E. L., & Kwok, V. (2016). Internet and smartphone application usage in eating disorders: A descriptive study in Singapore. *Asian Journal of Psychiatry, 19*, 50-55. doi:10.1016/j.ajp.2015.11.007

Taranis, L., Touyz, S., & Meyer, C. (2011). Disordered eating and exercise: development and preliminary validation of the compulsive exercise test (CET). *Eur Eat Disord Rev, 19*(3), 256-268. doi:10.1002/erv.1108

Teufel, M., Hofer, E., Junne, F., Sauer, H., Zipfel, S., & Giel, K. E. (2013). A comparative analysis of anorexia nervosa groups on Facebook. *Eating and Weight Disorders-Studies on Anorexia Bulimia and Obesity, 18*(4), 413-420. doi:10.1007/s40519-013-0050-y

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology, 63*(1), 163-173.

Walasek, L., Bhatia, S., & Brown, G. D. A. (2017). Positional goods and the social rank hypothesis: Income inequality affects online chatter about high and low status brands on Twitter. *Journal of Consumer Psychology*. doi:10.1016/j.jcps.2017.08.002

World Health Organization. (1993). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization.

1 **Tables**

2

3 **Table 1.** Names and descriptions of subreddits comprising the corpus

Subreddit name	Subreddit description†
Eating disorder subreddits	
r/proED	" **This is a place to discuss eating disorders, extreme/fringe eating behaviors, thinspo and recovery."
r/fuckeatingdisorders	" Eating disorders have many misconceptions, and part of that is because those who have it hide it since those who don't have it don't understand it because no one talks about it. FED is here to confront eating disorders and provide a place for anyone to ask questions."
r/EatingDisorders	" ## For Awareness, Information, and Questions about Recovering from EDs. We are a pro-recovery site, and only allow approved posts."
Health-related subreddits	
r/Fitness	"This subreddit is for discussion of physical fitness goals and how they can be achieved."
r/loseit	"A place for people of all sizes to discuss healthy and sustainable methods of weight loss. Whether you need to lose 2 lbs or 200 lbs, you are welcome here!"
r/nutrition	"A place to discuss eating well."

†*Descriptions of subreddits were extracted verbatim from reddit.com on December 31 2017*

4

5

1 **Table 2.** Characteristics of the corpus

Subreddits	Number of threads	Number of comments	Number of unique commenters†	Average number of comments by each unique commenter†	
				Mean (*SD*)	Median (range)
Eating disorder					
r/proED	37,335	387,357	11,123	32 (104)	5 (1:2,722)
r/fuckeatingdisorders	1,911	10,637	1,809	6 (19)	2 (1:569)
r/EatingDisorders	964	4,991	1,774	3 (6)	1 (1:133)
Health-related					
r/Fitness	213,885	4,620,754	382,426	11 (121)	2 (1:47,835)
r/loseit	108,496	1,836,704	131,825	13 (123)	2 (1:21,160)
r/nutrition	14,685	184,243	22,847	8 (44)	2 (1:2,821)

†Excludes commenter '[deleted]'

2

Figures

Fig. 1. Procedural flowchart

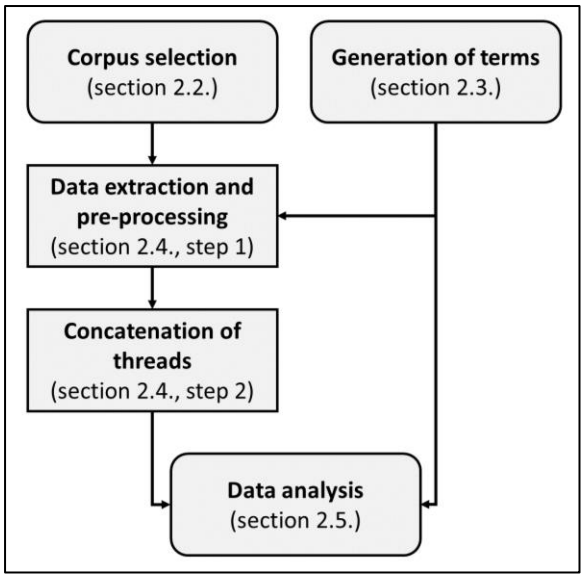

Fig. 2. Percentage of threads including at least one term from eating, body and exercise-related lists of terms

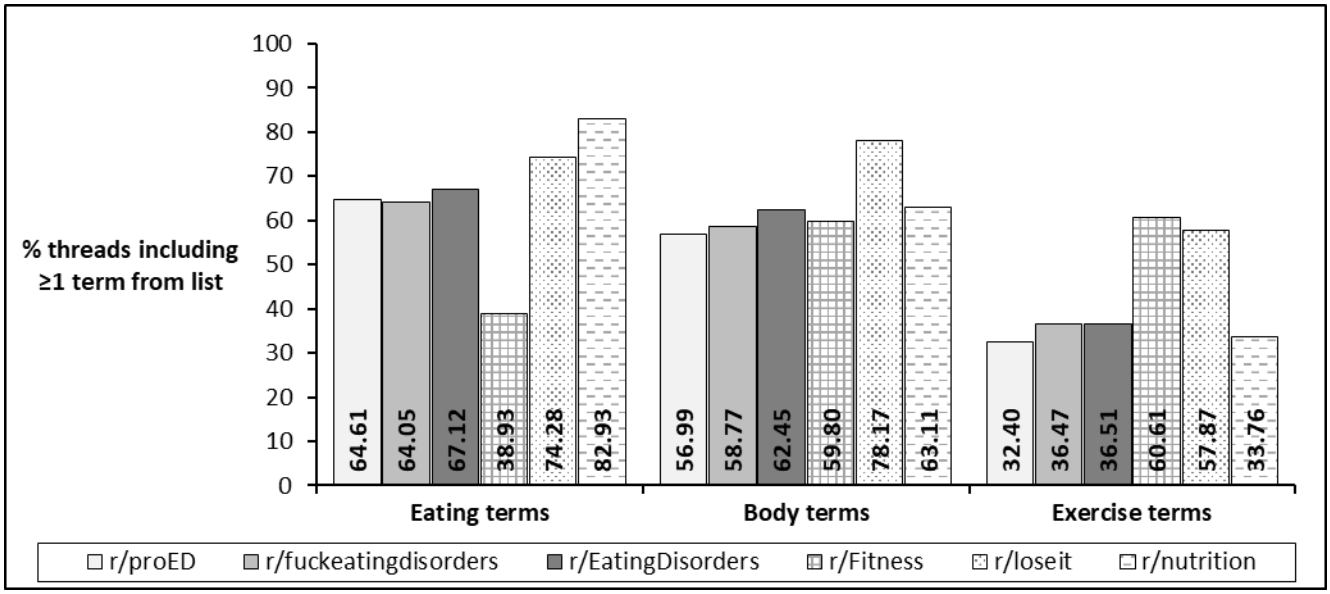

Fig. 3. Percentage of threads including at least one fitness tracker term

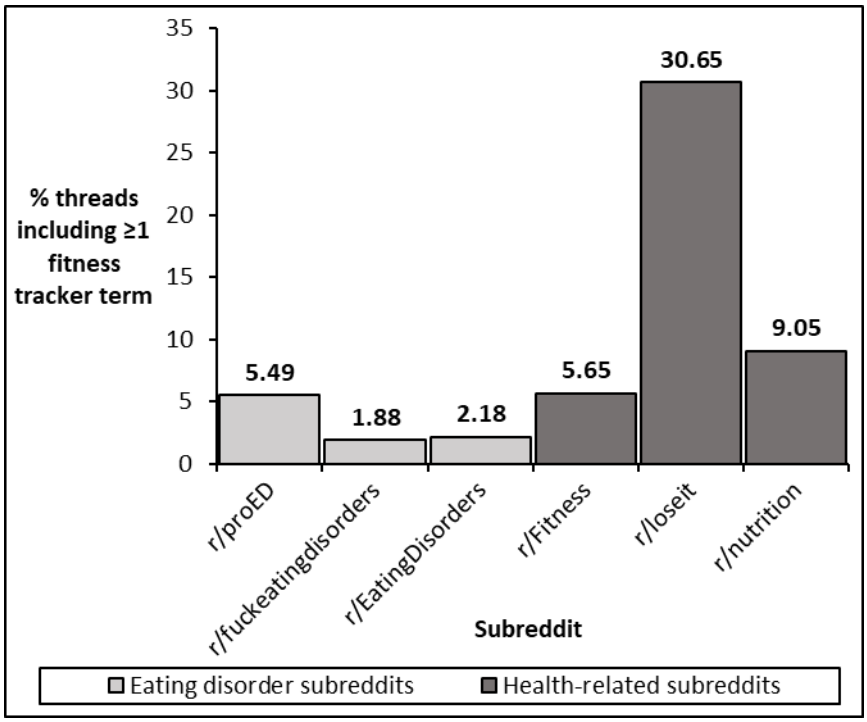

Fig. 4. Percentage of threads referencing fitness trackers that included at least one mention *of* *MyFitnessPal, Fitbit* and *Heart rate monitor*